Revista Brasileira de Computação Aplicada, Novembro, 2023

DOI: 10.5335/rbca.v15i3.14810
Vol. 15, Nº 3, pp. 15–24
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Métodos de classificação no espaço tangente na análise estatística de forma

Classification methods in tangent space in statistical shape analysis

Ariane Hayana Thomé de Farias [1] and Jhonnata Bezerra de Carvalho [1]

[1]Universidade Federal do Amazonas

*ariane.hayana@gmail.com; jhoncarvalho@ufam.edu.br

Recebido: 28/04/2023. **Revisado:** 05/07/2023. **Aceito:** 17/10/2023.

Resumo

O presente trabalho tem como propósito avaliar o desempenho de alguns classificadores da literatura em dados no espaço tangente no contexto da análise estatística de formas. Ademais, foram realizadas simulações considerando os cenários: (1) dados sem uso de *principal component analysis* (PCA); e (2) com uso de PCA utilizando as componentes que explicam de 70% a 75% e de 90% a 95% da variação total. Constatou-se na simulação que, quando há baixa concentração nos dados, o desempenho dos classificadores diminui, com ganhos expressivos na acurácia quando se fez o uso de PCA na maioria dos cenários observados. A etapa seguinte consistiu em realizar a classificação utilizando quatro aplicações em dados reais, considerando os mesmos cenários do estudo de simulação. Nestes, os melhores resultados foram observados em bancos de dados cujas formas médias eram expressivamente distintas entre os grupos. Por outro lado, os piores desempenhos foram observados em dados relacionados a ressonâncias magnéticas de pacientes esquizofrênicos, com acurácia máxima de 85,7%.

Palavras-Chave: Aprendizado de máquina; Coordenadas de Kendall; Coordenadas tangentes.

Abstract

The purpose of this study was to evaluate the performance of several classifiers from the literature on data in the tangent space at the statistical shape analysis context . Additionally, simulations were conducted considering the scenarios: (1) data without the principal component analysis (PCA) application ; and (2) data with the PCA application, where the components explained variations from 70% to 75% and from 90% to 95%. Simulation results showed the performance of the classifiers decreases when there is low concentration data, with significant accuracy gains observed when applying PCA in most of the scenarios examined. The next step was to perform classification using four datasets of real data, considering the same scenarios as in the simulation study. In these applications, the best results were observed in databases where the average shapes were significantly different between the groups. Conversely, the worst performances were observed in data related to magnetic resonance imaging of schizophrenic patients, with a maximum accuracy of 85.7%.

Keywords: Machine learning; Kendall coordinates; Tangent coordinates.

1 Introdução

A identificação de padrões em imagens tem sido amplamente utilizada nas mais diversas áreas do conhecimento e isso se deve especialmente ao avanço tecnológico que potencializou a necessidade de aplicações cada vez mais eficientes para o processamento digital de imagens. Com o avanço das ferramentas computacionais, a análise estatística de forma (*Statistical Shape Analysis*) tem possibilitado o estudo das formas geométricas de objetos, com aplicações em áreas como medicina, biologia, arqueologia, geografia, bem como em procedimentos de reconhecimento facial e diagnóstico de doenças por meio de ressonâncias magnéticas, entre outras áreas, conforme citam Dryden and Mardia (2016).

Alguns pesquisadores, ao longo dos anos, vêm aplicando métodos de classificação em dados de formas, como em Southworth et al. (2000) que utilizou um classificador baseado em redes neurais; Garcia et al. (2023) usaram *support vector machine* (SVM) e KNN (*k Nearest Neighbours*) para objetos em m-dimensões; Rodríguez et al. (2020) empregaram os métodos KNN, Fisher's linear discriminant analysis (FLDA) e o discriminante quadrático de Fischer combinando as divergências estocásticas de Rényi e Kullback-Leibler; Carvalho and Amaral (2020) aplicaram a SVM, FLDA e classificadores baseados em estatísticas de teste para dados no espaço das pré-formas e combinaram os classificadores por meio do método *ensemble*; Al-Dulaimi et al. (2019) usaram a SVM e FLDA para a classificação de células; e Karaci et al. (2020) empregaram classificadores como SVM, FLDA e também utilizaram métodos *ensemble*.

Ao utilizarmos dados relacionados às formas e tamanhos de objetos, é possível realizar a análise estatística destas formas, bem como a aplicação de aprendizado de máquina (*machine learning*) para a classificação, possibilitando, inclusive, a utilização de técnicas multivariadas.

Para tanto, Dryden and Mardia (2016) abordam alguns conceitos formais sobre o assunto para a compreensão da análise estatística de formas, a saber: a definição de *forma*, como um conjunto finito de pontos que são denominados de marcos (*landmarks*), que por sua vez, são fundamentais para a captura de pontos de correspondência em cada objeto e são utilizados para comparar o tamanho e a forma geométrica em estudo. Assim, faz-se relevante trazer a forma das imagens para o plano cartesiano e, neste trabalho, serão estudados objetos bidimensionais.

Os autores relatam também que existem três tipos de marcos:

- **Marcos anatômicos:** são pontos alocados por um especialista com a finalidade de identificar alguma característica científica específica da área em estudo. Assim, por exemplo, um especialista pode estar interessado em investigar características biológicas para a distinção do crânio de macacos machos e fêmeas através de marcos anatômicos em imagens.
- **Marcos matemáticos:** são pontos atribuídos considerando alguma propriedade matemática ou geométrica da imagem. Pode-se citar como exemplo, o reconhecimento de padrões em manuscritos para classificar automaticamente a letra correspondente.
- **Pseudo-Marcos:** são pontos localizados ao redor do contorno ou entre marcos anatômicos ou matemáticos.

Desta forma, a partir da apresentação de tais definições, pode-se visualizar na prática como são aplicados tais conceitos, que são melhor explanados nas seções seguintes. Com essa abordagem, é possível extrair informações relevantes sobre objetos e classificá-los usando algoritmos de aprendizagem de máquina e métodos de análise multivariada.

O presente trabalho foi baseado no estudo feito por de Farias (2020) e apresenta aplicações no espaço tangente de métodos da análise estatística de formas, utilizando aprendizado de máquina para a classificação de objetos bidimensionais, com objetivos de: (1) identificar padrões intrínsecos através da classificação de objetos no contexto da análise estatística de formas; (2) utilizar algoritmos de classificação em dados simulados, bem como aplicá-los em quatro bancos de dados reais; (3) avaliar a performance dos modelos e comparar com metodologias de extração de características.

Este trabalho encontra-se organizado em cinco seções, dispostas da seguinte forma: além desta introdução, a Seção 1 aborda os pilares teóricos referentes a análise estatística de formas, contextualizando definições sobre sistemas de coordenadas, bem como a abordagem relacionada às coordenadas tangentes. Na Seção 2 são descritas as metodologias utilizadas no trabalho, a descrição dos dados, os principais métodos de classificação com representações gráficas e algumas abordagens matemáticas. Posteriormente, na Seção 3 apresenta-se os resultados numéricos obtidos nas simulações e quatro aplicações de dados reais utilizando classificadores de *machine learning*. Na Seção 4 estão as considerações finais e, por fim, as referências bibliográficas.

Segundo Dryden and Mardia (2016), a análise estatística de formas é uma área da estatística responsável pelo estudo das formas geométricas de objetos. Dentre os pilares conceituais da análise estatística de formas cabe destacar o trabalho realizado por Kendall (1977), no qual faz um resumo e introduz uma nova representação de formas de objetos no espaço complexo. Pioneiro nesta temática, apresentou as primeiras formalizações e conceitos em seu trabalho intitulado "*Shape manifolds, procrustean metrics, and complex projective spaces*"(Kendall, 1984). Neste trabalho, o autor propõe um sistema de coordenadas atualmente conhecido como coordenadas de Kendall, no qual faz a utilização de pontos no plano cartesiano para obter a forma de um objeto. Assim, para exemplificar, a Fig. 1 ilustra a identificação de pontos em uma representação de uma estrela.

Neste exemplo, o objeto é identificado por um conjunto de pontos que representam o objeto no sistema de coordenadas. Assim como a estrela, outros objetos também podem ser facilmente representados e que podem passar por um processamento relacionado à análise estatística de formas. Tais processos envolvem a *forma, o tamanho e forma e a pré-forma*. Cada um desses difere entre si, de modo que:

- **Forma:** é toda a informação geométrica do objeto que resta quando são retirados os efeitos de localização, rotação e escala;

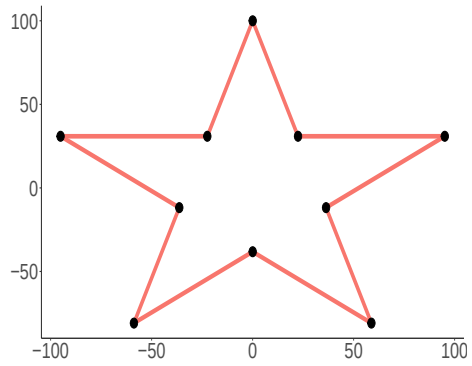


Figura 1: Representação de uma estrela com marcos matemáticos.

- **Tamanho e forma:** as transformações referem-se à remoção dos efeitos de localização e escala;
- **Pré-forma:** é definido pelas informações geométricas que permanecem no objeto retirando-se os efeitos de localização e escala.

Cada processo é composto por um conjunto de transformações que formam um sistema de coordenadas. Deste modo, define-se um $\widetilde{\mathbf{Y}}$ como sendo uma matriz com k marcos e m dimensões, ou seja, uma matriz de dimensão $k \times m$. O conjunto de marcos em um objeto é denominado *configuração*, portanto, o espaço de configuração será o espaço com todas as coordenadas dos marcos e $\widetilde{\mathbf{Y}}$ é definida como uma *matriz de configuração* (Dryden and Mardia, 2016), conforme apresentado na Eq. (1).

$$\widetilde{\mathbf{Y}} = \begin{pmatrix} \tilde{y}_{1,1} & \cdots & \tilde{y}_{1,m} \\ \vdots & \ddots & \vdots \\ \tilde{y}_{k,1} & \cdots & \tilde{y}_{k,m} \end{pmatrix}. \tag{1}$$

Neste estudo, os objetos analisados são bidimensionais, o que implica dizer que $m = 2$, que corresponde às formas planas. Com isso, a matriz de configuração é representada por

$$\widetilde{\mathbf{Y}} = \begin{pmatrix} \tilde{y}_{1,1} & \tilde{y}_{1,2} \\ \tilde{y}_{2,1} & \tilde{y}_{2,2} \\ \vdots & \vdots \\ \tilde{y}_{k,1} & \tilde{y}_{k,2} \end{pmatrix}.$$

Como mencionado anteriormente, para a obtenção da *forma* de um objeto, retira-se os efeitos de localização, rotação e escala, e, para tanto, a primeira transformação consiste em remover os efeitos de localização. Como estamos considerando uma matriz de configuração com duas dimensões ($m = 2$), esta pode ser reescrita como um vetor complexo $k \times 1$ tal que

$$\mathbf{z^0} = \left(\tilde{y}_{1,1} + i\tilde{y}_{1,2}, \ldots, \tilde{y}_{k,1} + i\tilde{y}_{k,2}\right)^T = \begin{pmatrix} \tilde{y}_{1,1} + i\tilde{y}_{1,2} \\ \tilde{y}_{2,1} + i\tilde{y}_{2,2} \\ \vdots \\ \tilde{y}_{k,1} + i\tilde{y}_{k,2} \end{pmatrix},$$

em que os elementos correspondem às coordenadas complexas dos marcos com a unidade imaginária $i = \sqrt{-1}$ e o símbolo "$\mathbf{0}$"(de $\mathbf{z^0}$) indica que a configuração conserva os efeitos de localização, escala e rotação (de Assis, 2018).

Para obter as coordenadas de Kendall (1984), precisamos primeiramente definir a submatriz Helmert ($\mathbf{H}$). A matriz de Helmert completa ($\mathbf{H}^F$) é uma matriz ortogonal com dimensão $k \times k$ em que a primeira linha possui todos os elementos iguais a $1/\sqrt{k}$ e a $(j+1)$-ésima linha para $j \geq 1$ dada por

$$\mathbf{h}_{j+1} = \left(h_j, \ldots, h_j, -jh_j, 0, \ldots, 0\right)^T, \quad h_j = -[j(j+1)]^{-1/2},$$

com $j = 1, \ldots, k-1$, no qual o número de zeros na linha $(j+1)$ é igual a $k - j - 1$ e a submatriz será dada por

$$\mathbf{H} = (\mathbf{h}_2, \ldots, \mathbf{h}_k)^T.$$

Nesta etapa, é possível remover o efeito de localização multiplicando-se o vetor complexo ($\mathbf{z^0}$) pela submatriz de Helmert ($\mathbf{H}$), o que resulta em

$$\omega = \mathbf{Hz^0}, \tag{2}$$

no qual ω representa a configuração sem o efeito de localização. A segunda transformação consiste na remoção do efeito de escala para encontrar a pré-forma através de um vetor complexo com dimensão $(k-1) \times 1$, assim, apenas a informação de rotação permanecerá. Para tal transformação, a escala é removida utilizando a configuração Helmertizada (ω) apresentada na Eq. (2), calculada da seguinte forma:

$$\mathbf{z} = \frac{\omega}{||\,\omega\,||} = \frac{\omega}{\sqrt{\omega^*\omega}} = \frac{\mathbf{Hz^0}}{\sqrt{(\mathbf{Hz^0})^*\mathbf{Hz^0}}}, \tag{3}$$

em que ω^* é um vetor transposto conjugado complexo de ω e $||.||$ é a norma complexa do vetor ω. Em seu trabalho, Kendall (1984) denota $\mathbf{z}$ como sendo a pré-forma da configuração complexa $\mathbf{z}^0$.

Ao obtermos um vetor complexo de dimensão $(k-1) \times 1$, o espaço das pré-formas será uma hiperesfera unitária complexa definida por

$$\mathbb{C}S^{k-2} = \{\mathbf{z} : ||\mathbf{z}|| = 1, \mathbf{z} \in \mathbb{C}^{k-1}\},$$

para $\mathbb{C}^{k-1}$ como sendo o espaço dos números complexos com dimensão $(k-1)$.

Por fim, obtemos o espaço da forma, que é formado pelo espaço da pré-forma sem o efeito de rotação. Para o espaço

da forma, este pode ser representado por

$$[\mathbf{z}] = \left\{ e^{i\theta}\mathbf{z} : \theta \in [0, 2\pi) \text{ e } \mathbf{z} \in \right\},$$

o qual denotamos por $[\mathbf{z}]$ qualquer versão rotacionada de $\mathbf{z}$ (pré-forma) e θ é o grupo especial ortogonal de rotações (de Oliveira, 2016).

Dryden and Mardia (2016) definem as *tangent coordinates* (TC) como uma aproximação linear do espaço de formas. Assim, considere uma amostra de pré-formas $\mathbf{z}_1, \ldots, \mathbf{z}_n$, em que as TC são dadas por

$$\mathbf{t}_j = e^{i\widehat{\theta}_j}[\mathbf{I}_{k-1} - \boldsymbol{\mu}\boldsymbol{\mu}^*]\mathbf{z}_j,\ j = 1, \ldots, n,$$

em que $\mathbf{I}_{k-1}$ é a matriz identidade $(k-1) \times (k-1)$, $\boldsymbol{\mu}$ é a forma média e $\widehat{\theta}_j = \arg(-\boldsymbol{\mu}^*\mathbf{z})$, em que a função arg(.) representa o argumento de um número complexo.

Usualmente utiliza-se o polo tangente como sendo a forma média obtida do conjunto de dados. Considerando que o espaço de *pré-formas* e *formas* são espaços não euclidianos, a análise estatística padrão é mais complexa (da Silveira, 2008), tornando-se mais vantajosa a utilização do espaço tangente.

Dentre as vantagens de utilizar as TC, destaca-se a possibilidade de utilizar diversos procedimentos de análise multivariada linear padrão, como por exemplo, avaliar a variabilidade da forma através de ferramentas como *Principal Component Analysis* (PCA) ou independent component analysis (ICA) aplicadas às TC.

Desta forma, podemos representar a visão geométrica das TC na esfera real, conforme Fig. 2.

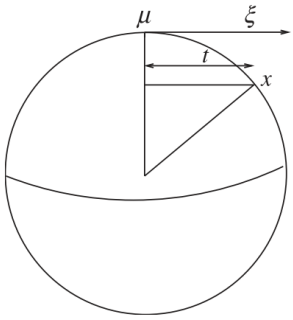


Figura 2: Visão geométrica das TC na esfera real (Dryden and Mardia, 2016).

2 Metodologia

Este trabalho está dividido em duas etapas, que consistem em realizar experimentos de simulação utilizando cinco métodos de classificação: SVM, FLDA, *Decision tree* (DT), *Random forest* (RF) e *Extreme gradient boosting* (XGBoost), os quais foram aplicados diversas configurações. Esses classificadores são amplamente utilizados na literatura em diversas áreas como: a SVM na classificação de dados de eletroencefalograma (Carvalho et al., 2019), na identificação do gosto musical de indivíduos (Lemos et al., 2019) e na predição do posicionamento de um jogador de futebol (Gasparini and Álvaro, 2017); DT na análise de dados faltantes (de Melo Júnior et al., 2020), para a predição da efetividade da substituição no futebol (Brutti et al., 2021) e na descoberta de conhecimentos em um jogo do tipo simulador (Dutra Júnior et al., 2021); FLDA usado na classificação de tumores em dados de expressão gênica (Dudoit et al., 2002); e o XGBoost empregado na predição de valores de moedas virtuais (Santos and de Paula, 2020).

Assim, na segunda etapa foram utilizadas quatro bases de dados reais distintas com o objetivo de analisar o desempenho dos classificadores. Nas aplicações da segunda etapa utilizou-se o método *leave-one-out* na validação cruzada, tendo em vista que o formato dos dados em estudos possibilitou a adoção de tal etapa. Por outro lado, nas simulações a quantidade de dados era maior, sendo mais adequado o método de validação cruzada *k-fold*.

Os conjuntos de dados reais utilizado neste estudo são oriundos do pacote SHAPES sob autoria de Dryden (2021). Este pacote inclui diversos métodos e bancos de dados reais para a realização da análise estatística de formas em R, com explicações detalhadas disponíveis na obra de Dryden and Mardia (2016).

2.1 Classificação

Neste trabalho foram utilizados diversos classificadores, a SVM proposta por Boser et al. (1992) e Vapnik (1995); FLDA proposta por Fisher no século XX (Hastie et al., 2009) e, modelos baseados em árvores: DT, RF e XGBoost. Segundo Morettin and Motta (2022), modelos baseados em árvores foram desenvolvidos por Leo Breiman e são bastante populares devido à simplicidade em termos conceituais e computacionais e no decorrer dos anos, foram amplamente exploradas, servindo de base para generalizações tais como RF e XGBoost. Este último, por exemplo, foi apresentado pelos autores Chen and Guestrin (2016), que criaram uma implementação do algoritmo *Gradient Boosting* (propostos por Friedman et al. (2000)). A motivação para a criação deste novo algoritmo consistia, especialmente, na utilização de menos recursos em um sistema com melhor desempenho em comparação aos demais, sendo este um sistema de aumento de árvore escalável, reduzindo o tempo de treinamento ao lidar com grandes conjuntos de dados.

Na comparação entre os classificadores, os resultados obtidos na matriz de confusão foram avaliados e obteve-se as acurácias por classificador considerando três cenários distintos, sendo que no primeiro momento utilizou-se a totalidade dos dados disponíveis, no segundo momento fez-se a redução de dimensionalidade através da técnica de PCA, que proporcionou uma transformação nos dados originais, que foram projetados e centrados no plano tangente tendo como polo a forma média. Nesta etapa fez-se a seleção das primeiras componentes que explicavam entre 70% a 75%, e, no último cenário considerou-se o intervalo de 90% a 95%. O Algoritmo 1 resume os passos para a aplicação da PCA e treinamento dos classificadores.

Algoritmo 1: Utilização da PCA nas amostras de objetos.

1 Sejam $\tilde{\mathbf{Y}}_{i1}, \tilde{\mathbf{Y}}_{i2}, \ldots, \tilde{\mathbf{Y}}_{in_i}$ uma amostra de objetos de tamanho n_i para o i-ésimo grupo, com $i = 1, 2$.
2 Considere $\mathbf{t}_{ji}$ como sendo as TC para o j-ésimo objeto do i-ésimo grupo, em que o polo tangente é a estimativa da forma média μ considerando todos os objetos.
3 A PCA é aplicada em cada amostra de coordenadas tangentes. Escolhe o número n_c de fatores, tal que o percentual de variação explicada fique em torno de 70% a 75% (70%-75%) ou de 90% a 95% (90%-95%).
4 Utilizar os escores de cada grupo no classificador.

2.2 Estudo de simulação

Com o objetivo de avaliar o desempenho dos classificadores, foi conduzido um estudo de simulação com dados sintéticos no qual a Fig. 1 foi replicada em duas posições distintas formando dois grupos diferentes. As coordenadas bidimensionais de 10 marcos presentes em cada figura foram dispostas de acordo com a Fig. 3, em que o *Grupo 1* refere-se a estrela em laranja e o *Grupo 2* a estrela em azul. A etapa de simulação teve como finalidade a comparação entre os classificadores utilizados e a criação de um cenário controlado com um quantitativo maior de marcos (na literatura, estudos de simulação foram realizados também por Carvalho (2019), Garcia et al. (2023) e Rodríguez (2015)), além de contribuir para extrair informações que auxiliaram na tomada de decisão sobre quais estratégias de ajuste dos parâmetros seria a mais adequada para avaliar o desempenho em diferentes cenários.

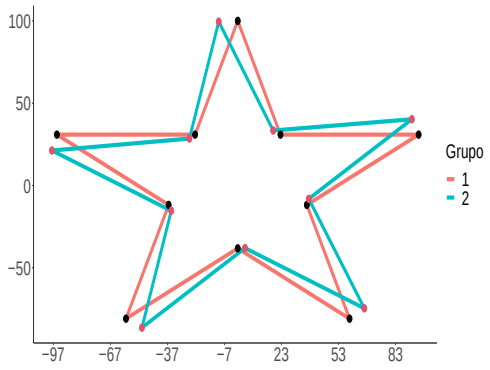


Figura 3: Formas médias das estrelas.

Para avaliar os classificadores, foram utilizados os mesmos modelos mencionados anteriormente na Subseção 2.1. Em cada classificador foram aplicadas configurações distintas, conforme os parâmetros utilizados em cada algoritmo e para cada um destes foram calculadas as médias das acurácias para identificar quanto que o modelo está classificando corretamente.

As configurações utilizadas para cada classificador foram:

- SVM: o classificador SVM está implementado no pacote `kernlab`. Utilizou-se três tipos de *kernel*, sendo estes o linear, polinomial e gaussiano com os custos C = 1, 10, 100, 1.000, 10.000 e 100.000. No *kernel* polinomial, os graus foram variados de 1 a 5, sendo que o grau 1 corresponde ao *kernel* linear. Quanto ao *kernel* gaussiano, a configuração adotada (além do ajuste de C), foi o ajuste no parâmetro σ_G (largura de banda da função *kernel*), com parâmetros iguais a σ_G = 0,01; 0,05; 0,1; 0,5; 1; 100; 1.000; 10.000 e 100.000;
- XGBoost: o classificador XGBoost está implementado no pacote `xgboost`. No qual existem duas versões: linear (`gblinear`) e árvore (`gbtree`). No presente trabalho foi utilizada a versão `gbtree`, no qual foram usadas as seguintes configurações de parâmetros: η = 0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9 e 1 (encolhimento do tamanho do passo usado na atualização para evitar o *overfitting*, $\eta \in [0, 1]$); γ = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10 (redução mínima de perda necessária para fazer uma nova partição em um nó folha da árvore, $\gamma \in [0, \infty]$), `max_depth` = 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10 (profundidade máxima de uma árvore, `max_depth` $\in [0, \infty]$) e `nrounds` = 2, 4, 6, 8, 10 e 12 (número interações de reforço no treinamento);
- FLDA: foi implementado pelos autores e não possui configuração de parâmetros;
- DT: o classificador DT está implementado no pacote `rpart` e não possui configurações de parâmetros; e
- RF: o classificador RF está implementado no pacote `randomForest` e foi utilizado parâmetro `ntree` (número de árvores) variando de 3 a 500 árvores.

Tais configurações foram utilizadas para encontrar uma combinação eficiente em diferentes cenários e avaliar a generalização do modelo.

A validação cruzada adotada foi a *k-fold*, com n = 100 amostras divididas em k = 5 partições, ou seja, para cada grupo foram utilizados 100 elementos da amostra, os quais foram distribuídos em 5 subconjuntos, os quais foram separados em treino (80%) e teste (20%) em cada grupo.

Com o intuito de realizar a classificação no espaço tangente, adotou-se o procedimento de separação e projeção dos dados simulados no plano tangente, tendo como referência a forma média.

Os testes de classificação foram elaborados com uma distribuição Watson complexa, embasados nos procedimentos adotados por Carvalho (2019), de modo que foram realizadas diferentes configurações de kapa (κ), variando o grau de dispersão dos dados em torno do centro, em que quanto maior o valor de κ maior a concentração (menor variância), por outro lado, quanto menor o valor de κ maior a dispersão (maior variância) (Dryden and Mardia, 2016). Desta forma, as configurações do parâmetro κ foram: 10.000, 9.000, 8.000, 7.000, 6.000, 5.000, 4.000, 3.000 e 2.000.

A seguir, comparou-se os resultados obtidos em cada classificador considerando três cenários:

- Classificação sem PCA: utilizou-se integralmente o conjunto de dados original em cada grupo;
- Classificação com 70% − 75%: neste, considerou-se as

Tabela 1: Acurácia para os classificadores utilizando o método *k-fold* para diferentes valores de κ.

Classificador		κ								
		10.000	9.000	8.000	7.000	6.000	5.000	4.000	3.000	2.000
Sem PCA		0,930	**0,960**	0,900	0,930	0,895	0,855	0,880	0,835	0,695
PCA: 70% – 75%	SVM Linear	**0,940**	0,945	0,920	0,950	**0,910**	**0,895**	**0,920**	**0,840**	**0,920**
PCA: 90% – 95%		0,935	0,950	**0,925**	**0,955**	0,885	0,880	0,890	**0,840**	0,890
Sem PCA		**0,935**	**0,930**	**0,905**	**0,910**	**0,880**	**0,820**	**0,860**	**0,770**	0,645
PCA: 70% – 75%	SVM Polinomial	0,925	0,860	0,870	0,885	0,830	0,805	0,805	0,745	**0,805**
PCA: 90% – 95%		0,900	0,885	0,860	0,895	0,850	0,795	0,805	0,725	**0,805**
Sem PCA		**0,940**	**0,955**	**0,950**	**0,950**	**0,885**	**0,900**	0,870	**0,850**	**0,730**
PCA: 70% – 75%	SVM Gauss	0,925	0,935	0,945	**0,950**	0,870	**0,900**	**0,875**	0,830	**0,730**
PCA: 90% – 95%		0,925	0,935	0,945	**0,950**	0,870	**0,900**	**0,875**	0,830	**0,730**
Sem PCA		**0,940**	0,940	0,925	**0,950**	**0,880**	0,870	0,890	0,845	0,700
PCA: 70% – 75%	FLDA	**0,940**	**0,950**	**0,940**	**0,950**	0,870	**0,900**	**0,915**	**0,850**	**0,760**
PCA: 90% – 95%		0,935	**0,960**	0,935	0,940	0,875	0,870	0,895	0,835	0,720
Sem PCA		0,890	0,800	0,825	0,875	0,750	0,750	0,785	0,755	0,640
PCA: 70% – 75%	DT	**0,915**	**0,950**	**0,950**	**0,925**	0,870	**0,860**	**0,870**	0,765	0,710
PCA: 90% – 95%		**0,915**	**0,950**	**0,950**	**0,925**	**0,885**	**0,860**	**0,870**	**0,775**	**0,720**
Sem PCA		**0,970**	0,965	0,960	**0,975**	0,925	**0,950**	0,895	**0,885**	0,795
PCA: 70% – 75%	RF	0,960	0,975	0,965	0,970	0,925	0,920	**0,930**	0,865	0,815
PCA: 90% – 95%		0,965	**0,980**	**0,975**	0,970	**0,930**	0,920	**0,930**	0,860	**0,820**
Sem PCA		0,940	**0,950**	0,910	0,915	0,900	0,885	0,870	0,815	0,750
PCA: 70% – 75%	*XGBoost*	**0,960**	0,940	**0,935**	0,930	0,930	**0,895**	0,875	0,840	0,755
PCA: 90% – 95%		0,950	0,940	0,935	**0,935**	**0,940**	0,885	**0,885**	**0,845**	**0,775**

combinações lineares dos atributos originais, de modo que foram selecionadas apenas as primeiras componentes principais que explicavam entre 70% a 75% a variação dos dados; e
- Classificação com 90% – 95%: realizou-se a redução de dimensionalidade dos dados considerando apenas as primeiras componentes principais que explicavam entre 90% a 95% a variação dos dados, mantendo-se somente as informações mais importantes no intervalo mencionado.

A Tabela 1 possui as acurácias para os classificadores utilizados no estudo considerando os três cenários do uso da PCA. De forma geral para todos os classificadores, há uma tendência decrescente para os valores das acurácias, ou seja, na medida que o valor de κ diminui a acurácia tende a diminuir. Essa tendência é esperada, pois quanto menor o valor de κ maior a variabilidade dados em torno da forma média, o que acarreta em uma tarefa mais difícil para classificar corretamente.

- **SVM**: em relação à SVM linear, pode-se perceber na maioria das situações a acurácia do classificador possui uma melhora quando é combinado à PCA. Por outro lado, em todos os cenários a SVM com *kernel* polinomial combinada com a PCA, não forneceu uma melhora na acurácia, exceto para o caso em que κ = 2.000. Em relação à SVM com *kernel* gaussiano, nota-se um desempenho muito parecido entre as acurácias, no qual em muitas situações os resultados foram idênticos.
- **FLDA**: observa-se que na maior parte dos cenários, o classificador FLDA combinado com a PCA forneceu as melhores acurácias, exceto para os cenários com κ = 10.000 e 7.000, nos quais houve empate, e para κ = 6.000 em que o FLDA sem PCA forneceu o melhor resultado.
- **DT**: para o classificador árvore de decisão, pode-se notar que em todos os cenários a acurácia foi melhor quando se é utilizado classificador combinado com a PCA.
- **RF**: pode-se constatar que o classificador floresta aleatória combinado com a PCA forneceu as melhores acurácias em comparação com o mesmo classificador sem o uso da PCA, exceto para os cenários em que κ = 10.000, 7.000, 5.000 e 3.000.
- ***XGBoost***: nota-se que o uso da PCA melhorou a acurácia do classificador em todos os cenários, exceto para o cenário no qual κ = 9.000.

De maneira geral, pode-se notar que em muitas situações há um ganho quando se é utilizado a PCA no classificador. Além disso, pode-se ver que os piores desempenhos acontecem sem o uso da PCA para os classificadores DT e *XGBoost*.

2.3 Aplicações

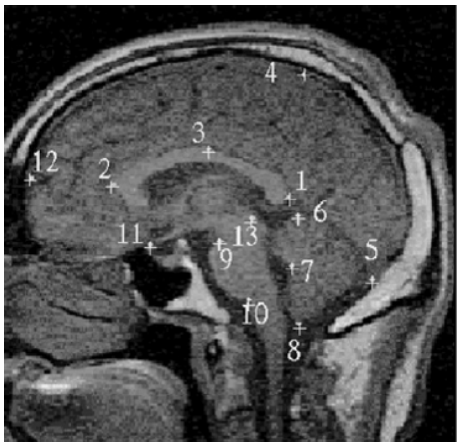
As aplicações utilizadas neste estudo foram selecionadas considerando coordenadas cartesianas bidimensionais, cujas descrições e formatos das matrizes de configuração são apresentados na Tabela 2.

Tabela 2: Resumo das características dos bancos de dados.

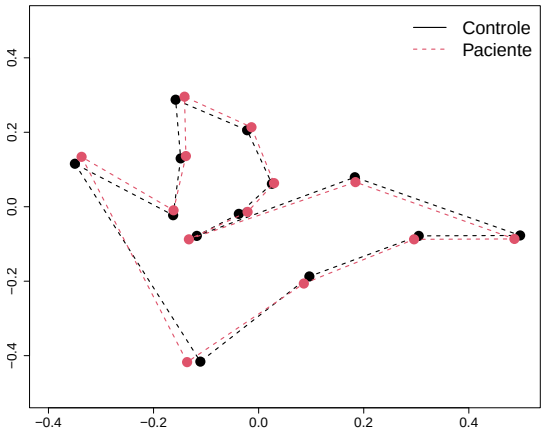
Dados	Formato	Grupos
Esquizofrenia	13 × 2 × 28	Controle e paciente
Crânios de Gorilas	8 × 2 × 59	Fêmea e macho
Vértebras de camundongos	6 × 2 × 46	Grande e pequeno
Crânios de ratos	8 × 2 × 36	Dias 7 e 150

Fonte: Elaboração própria dos autores.

A descrição de cada banco de dados é apresentada nas subseções seguintes.



(a) Imagem de uma ressonância magnética cerebral. Fonte: Dryden and Mardia (2016).



(b) As formas médias do grupo *controle* e *paciente*.

Figura 4: Representação do objeto e formas médias dos grupos.

2.3.1 *Esquizofrenia*

Consiste em dados cerebrais coletados de ressonâncias magnéticas de voluntários e pacientes esquizofrênicos com intuito de identificar quaisquer diferenças entre o grupo de controle e o grupo de pacientes com esquizofrenia. O estudo considera dois grupos, que são compostos por 14 voluntários do grupo controle e 14 pacientes com esquizofrenia, ou seja, 28 pacientes ao todo, considerando 13 marcos para cada um destes, de acordo com a Fig. 4.

2.3.2 *Crânios de primatas grandes - Gorilas*

Neste, utilizou-se dados cranianos de macacos grandes de modo a avaliar as diferenças entre crânios por sexo. Para esta análise, optou-se pela investigação das diferenças na espécie *Gorila*. Assim, analisou-se a amostra correspondente a 30 gorilas fêmeas e 29 gorilas machos na fase adulta. Para cada grupo, observou-se 8 marcos, que podem ser expressos graficamente através de suas formas médias de acordo com a Fig. 5.

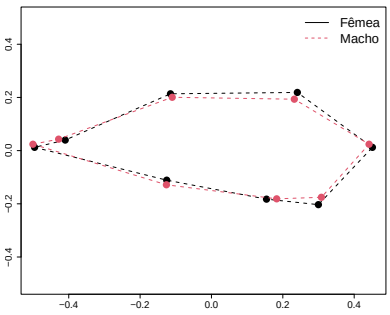


Figura 5: Formas médias de gorilas machos e fêmeas.

2.3.3 *Vértebras de camundongos*

Este conjunto de dados é oriundo de um experimento que consistia em avaliar os efeitos do peso corporal sobre a forma das vértebras de camundongos. Para conduzir o experimento, os camundongos foram divididos em três grupos: *grande*, *pequeno* e do grupo *controle*. A separação entre os grupos *grande* e *pequeno* foi feita com base no peso corporal dos animais, sendo que o grupo *grande* incluiu os camundongos com maior peso a cada geração, enquanto o grupo *pequeno* incluiu aqueles com o menor peso. O terceiro grupo, controle, foi composto por camundongos que não foram selecionados.

No presente trabalho optou-se pela análise bidimensional, portanto, os grupos selecionados foram os com peso corporal *grande* e *pequeno*, que compõem a base de dados estruturada com coordenadas bidimensionais de 6 marcos para cada um dos 46 camundongos. A avaliação foi feita considerando as diferenças na forma e tamanho na segunda vértebra torácica T2 de 23 dos camundongos do grupo *grande* e 23 do grupo *pequeno*. Na Fig. 6 pode-se observar as formas médias dos camundongos para cada grupo.

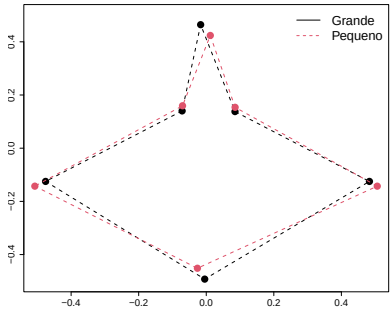


Figura 6: Formas médias das vértebras dos camundongos.

Tabela 3: Acurácia para os classificadores utilizando o método *leave-one-out*.

Classificador		Bases de dados			
		Camundongos	Esquizofrenia	Gorilas	Ratos
Sem PCA		0,935	0,679	0,983	**1,000**
PCA: 70% – 75%	SVM Linear	0,935	0,714	0,915	**1,000**
PCA: 90% – 95%		**0,957**	**0,786**	**1,000**	**1,000**
Sem PCA		0,870	0,643	**0,932**	**1,000**
PCA: 70% – 75%	SVM Polinomial	**0,891**	**0,714**	**0,932**	**1,000**
PCA: 90% – 95%		0,848	**0,714**	0,797	**1,000**
Sem PCA		0,913	0,714	**1,000**	**1,000**
PCA: 70% – 75%	SVM Gauss	0,935	**0,786**	0,915	**1,000**
PCA: 90% – 95%		**0,978**	0,714	**1,000**	**1,000**
Sem PCA		0,761	0,464	0,729	0,806
PCA: 70% – 75%	FLDA	**0,870**	0,679	**0,797**	**1,000**
PCA: 90% – 95%		0,848	**0,714**	**0,797**	**1,000**
Sem PCA		**0,957**	**0,786**	0,831	**1,000**
PCA: 70% – 75%	DT	0,826	0,679	**0,881**	**1,000**
PCA: 90% – 95%		0,826	0,679	**0,881**	**1,000**
Sem PCA		**0,957**	0,714	0,949	**1,000**
PCA: 70% – 75%	RF	0,891	**0,857**	0,949	**1,000**
PCA: 90% – 95%		0,891	**0,857**	**0,966**	**1,000**
Sem PCA		**0,957**	**0,786**	**0,932**	**1,000**
PCA: 70% – 75%	*XGBoost*	0,870	0,750	**0,932**	**1,000**
PCA: 90% – 95%		0,848	0,714	**0,932**	**1,000**

2.3.4 *Crânios de ratos*

Dados extraídos de raios-X cranianos de ratos com intuito de descrever as mudanças decorrentes do crescimento deles. Para tanto, foram registradas as coordenadas bidimensionais de radiografias dos crânios em diversas etapas do desenvolvimento de cada um dos 18 ratos em 7, 14, 21, 30, 40, 60, 90 e 150 dias de vida. Em cada radiografia foram identificados 8 marcos para cada idade de crescimento. Para este trabalho, com intuito de analisar as mudanças entre o início e o fim das etapas de crescimento, optou-se por avaliar os dias 7 e 150, os quais as formas médias são expressas graficamente na Fig. 7.

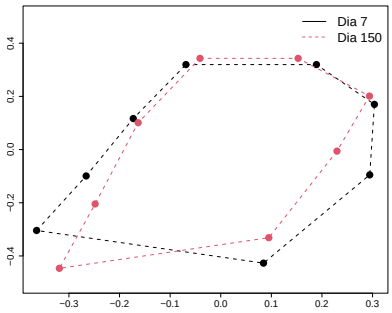


Figura 7: Formas médias de crânios de ratos.

3 Resultados e discussões

No presente estudo foi realizado um estudo de simulação e pôde-se perceber que à medida que a concentração diminui a acurácia diminui, ou seja, quanto maior a variabilidade menor a acurácia para todos os classificadores. Além disso, percebeu-se também que a PCA, de forma geral, melhora a acurácia dos classificadores.

Os algoritmos de classificação foram utilizados em quatro bases de dados. Os melhores desempenhos dos classificadores foram com a base de dados sobre crânios de ratos, o que era esperado, pois existe uma diferença expressiva entre as formas médias para os dois grupos (ver Fig. 7) e isso facilitou na classificação. Por outro lado, os classificadores reduziram as acurácias na base de dados de esquizofrenia, os melhores desempenhos foram obtidos pela SVM linear e a SVM com *kernel* gaussiano, ambos combinados com a PCA. Para os dados de camundongos e de gorilas, os desempenhos dos classificadores foram bem similares, exceto para o classificador FLDA que forneceu os piores resultados.

Na Tabela 3 temos as acurácias dos classificadores para as bases de dados consideradas.

- Camundongos: o classificador SVM com *kernel* Gaussiano combinado com a PCA forneceu a melhor acurácia, correspondendo a 97,8%; seguido pela SVM linear com PCA (90%-95%), XGBoost, DT e RT sem PCA, que atingiram uma acurácia de 95,7%;
- Esquizofrenia: o classificador RF atingiu a maior acurácia correspondendo a 85,7% com o uso da PCA. Além disso, note que as melhores acurácias para essa base de dados foram alcançadas quando o classificador é combinado com a PCA, exceto para os classificadores DT e XGBoost.
- Gorilas: para essa base de dados a SVM linear e combinada com a PCA forneceu uma acurácia de 100,0%. Ademais, a SVM com *kernel* gaussiano com e sem a PCA obtiveram o mesmo desempenho, correspondendo a uma taxa de 100,0% de acerto.
- Ratos: nessa aplicação pode-se observar que todos os classificadores atingiram uma acurácia de 100,0%, exceto para o classificador FLDA sem o uso da PCA.

4 Considerações finais

O presente trabalho tem como propósito avaliar o desempenho de alguns classificadores da literatura em dados no espaço tangente na análise estatística de formas. Ademais, foi utilizada a PCA no espaço tangente com o intuito de melhorar o desempenho dos classificadores.

Considerando os três cenários avaliados, fez-se, inicialmente, um estudo de simulação, no qual constatou-se que, quando há baixa concentração nos dados, o desempenho dos classificadores diminui, o que podemos inferir que o classificador apresentou resultados menos exitosos na classificação de novos dados. Em uma análise global, um melhor desempenho dos classificadores foi obtido quando combinados com PCA, cabendo destaque para a SVM, em que as acurácias foram majoritariamente altas em comparação ao cenário sem PCA.

Após o estudo de simulação, fez-se a avaliação considerando dados reais de quatro bancos de dados. Para tanto, utilizou-se os cenários adotados nas simulações em condições semelhantes na abordagem de configuração, diferindo somente na técnica de validação cruzada, que neste caso, foi a *leave-one-out*. Neste, elaborou-se preliminarmente visualizações gráficas comparativas entre os grupos para obter informações sobre a disposição dos marcos e suas respectivas formas médias. No banco de dados de Esquizofrenia, por exemplo, é notória a semelhança nas formas médias para o grupo controle e os indivíduos com esquizofrenia (ver Fig. 4). Tal proximidade entre as formas contribuiu para um baixo desempenho dos diversos classificadores utilizados no estudo, alcançando acurácia máxima de 85,7% com o classificador RF no cenário o qual houve redução de dimensionalidade com PCA (em ambos os cenários). Por outro lado, percebe-se que no banco de dados referente aos crânio de ratos, as diferenças nas formas médias ocorrem expressivamente entre o 7º dia de vida e o 150º dia (conforme apresentado na Fig. 7), e, neste caso, os classificadores apresentaram, em sua maioria, acurácia máxima de 100%, com exceção do classificador FLDA, com acurácia de 80,6% sem o uso de PCA.

Com relação aos demais bancos de dados, verificou-se que na aplicação com vértebras de camundongos, o desempenho dos classificadores foi acima de 76%, com o menor desempenho observado na FLDA sem uso de PCA. Na aplicação em dados cranianos de gorilas, assim como em camundongos, a menor acurácia ocorreu quando fez-se a classificação considerando FLDA sem PCA (com aproximadamente 73%). Assim, pode-se observar que, embora existam diferenças percentuais nos resultados obtidos entre classificadores, estes se mostram eficazes na categorização dos conjuntos de dados.

Ademais, sugere-se como trabalho futuro a utilização de outros classificadores para avaliar o desempenho dos algoritmos, assim como a utilização de técnicas como *data augmentation* para a criação de dados sintéticos. Trabalhos como os de Anaya-Isaza and Mera-Jiménez (2022) mostram, por exemplo, a utilização desta técnica na detecção de tumores cerebrais em imagens de ressonância magnética, sendo esta, uma promissora técnica para melhorar a capacidade de generalização do algoritmo.

Referências

Al-Dulaimi, K., Chandran, V., Nguyen, K., Banks, J. and Tomeo-Reyes, I. (2019). Benchmarking hep-2 specimen cells classification using linear discriminant analysis on higher order spectra features of cell shape, *Pattern Recognition Letters* **125**: 534–541. https://doi.org/10.1016/j.patrec.2019.06.020.

Anaya-Isaza, A. and Mera-Jiménez, L. (2022). Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging, *IEEE Access* **10**: 23217–23233. https://doi.org/10.1109/ACCESS.2022.3154061.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, New York, NY, USA, pp. 144–152. https://doi.org/10.1145/130385.130401.

Brutti, N. S., Duarte, D. and Dal Bianco, G. (2021). Predição da efetividade da substituição no futebol: caso campeonato brasileiro da série A, *Revista Brasileira de Computação Aplicada* **13**(1): 42–52. https://doi.org/10.5335/rbca.v13i1.11120.

Carvalho, J. B. (2019). *Métodos de classificação e bondade de ajuste na análise de formas planas*, Doutorado em estatística, Universidade Federal de Pernambuco, Recife. https://repositorio.ufpe.br/handle/123456789/35508.

Carvalho, J. B. and Amaral, G. J. A. (2020). Classification methods for planar shapes, *Expert Systems with Applications* **151**: 113320. https://doi.org/10.1016/j.eswa.2020.113320.

Carvalho, J. B., Silva, M. C., von Borries, G. F., de Pinho, A. L. S. and von Borries, R. F. (2019). A combined fourier analysis and support vector machine for eeg classification, *Chilean Journal of Statistics (ChJS)* **10**(2). https://soche.cl/chjs/volumes/10/ChJS-10-01-01.pdf.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

da Silveira, F. V. J. (2008). *Testes de permutação e bootstrap em análise estatística de formas: aplicações à zoologia*, Mestrado em estatística, Universidade Federal de Pernambuco, Recife. https://repositorio.ufpe.br/handle/123456789/6156.

de Assis, E. C. (2018). *Algoritmos de particionamento aplicados à análise estatística de formas*, Doutorado em ciência da computação, Universidade Federal de Pernambuco, Recife. https://repositorio.ufpe.br/handle/123456789/31431.

de Farias, A. H. T. (2020). *Análise estatística de formas: uma aplicação de técnicas de machine learning no espaço tangente*, Graduação em estatística, Universidade Federal Amazonas, Manaus.

de Melo Júnior, G., Alcalá, S. G. S., Furriel, G. P. and Vieira, S. L. (2020). Missing data analysis using machine learning methods to predict the performance of technical students, *Revista Brasileira de Computação Aplicada* **12**(2): 134–143. https://doi.org/10.5335/rbca.v12i2.10565.

de Oliveira, R. A. (2016). *Algoritmos para determinação do número de grupos em estudos de formas planas*, Mestrado em estatística, Universidade Federal de Pernambuco, Recife. https://repositorio.ufpe.br/handle/123456789/17314.

Dryden, I. L. (2021). *Shapes package*, R Foundation for Statistical Computing, Vienna, Austria. Contributed package, Version 1.2.6. http://www.R-project.org.

Dryden, I. L. and Mardia, K. V. (2016). *Statistical Shape Analysis with applications in R*, John Wiley & Sons.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American statistical association* **97**(457): 77–87. https://doi.org/10.1198/016214502753479248.

Dutra Júnior, N. C. d. S., Billa, C. Z. and Adamatti, D. F. (2021). Descoberta de conhecimento em um jogo sério para o ensino de plantas industriais: um estudo de caso utilizando árvores de decisão, *Revista Brasileira de Computação Aplicada* **13**(1): 98–111. https://doi.org/10.5335/rbca.v13i1.11378.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), *The annals of statistics* **28**(2): 337–407. https://doi.org/10.1214/aos/1016218223.

Garcia, V. G. i., Gual-Arnau, X., Ibáñez, M. V. and Simó, A. (2023). A gaussian kernel for kendall's space of md shapes, *Pattern Recognition* p. 109887. https://doi.org/10.1016/j.patcog.2023.109887.

Gasparini, R. and Álvaro, A. (2017). Análise entre algoritmos de aprendizado de máquina para suportar a predição do posicionamento do jogador de futebol, *Revista Brasileira de Computação Aplicada* **9**(2): 70–83. https://doi.org/10.5335/rbca.v9i2.6454.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer.

Karaci, A., Ozkaraca, O., Acar, E. and Demir, A. (2020). Prediction of traumatic pathology by classifying thorax trauma using a hybrid method for emergency services, *IET Signal Processing* **14**(10): 754–764. https://doi.org/10.1049/iet-spr.2020.0014.

Kendall, D. G. (1977). The diffusion of shape, *Advances in applied probability* **9**(3): 428–430. https://doi.org/10.2307/1426091.

Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces, *Bulletin of the London mathematical society* **16**(2): 81–121. https://doi.org/10.1112/blms/16.2.81.

Lemos, J. C., dos Santos, M. C. B., Vilela, P. R. S. and de Rezende, M. N. (2019). Aplicando aprendizado de máquina para identificação do gosto musical de um indivíduo, *Revista Brasileira de Computação Aplicada* **11**(3): 88–98. https://doi.org/10.5335/rbca.v11i3.9230.

Morettin, P. A. and Motta, S. J. d. (2022). *Estatística e ciência de dados*, LTC, Rio de Janeiro, RJ.

Rodríguez, W. D. A. (2015). *Classificação estatística nos espaços euclideanos e não-euclideanos com aplicação em dados de forma*, Mestrado em estatística, Universidade Federal de Pernambuco, Recife.

Rodríguez, W. D. A., Amaral, G. J. A., Nascimento, A. D. C. and Ferreira, J. A. (2020). Information criteria in classification: new divergence-based classifiers, *Journal of Statistical Computation and Simulation* **90**(17): 3200–3217. https://doi.org/10.1080/00949655.2020.1798445.

Santos, W. R. and de Paula, H. B. (2020). Predição de valores de moedas virtuais através da análise de sentimento de notícias e tweets, *Revista Brasileira de Computação Aplicada* **12**(1): 1–15. https://doi.org/10.5335/rbca.v12i1.8831.

Southworth, R., Mardia, K. V. and Taylor, C. C. (2000). Transformation-and label-invariant neural network for the classification of landmark data, *Journal of Applied Statistics* **27**(2): 205–215. https://doi.org/10.1080/02664760021745.

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer, New York.